图书情报工作
第63卷 第16期 2019年8月

热点论文分布特征与影响因素分析*
——兼评时间窗口与学科间差异

宋超 陈悦 汪玲 左佳 刘则渊
大连理工大学科学学与科技管理研究所暨WISE实验室 大连 116024

摘要：［目的/意义］近年来，热点论文逐渐受到学术界重视，为数不多的研究成果已开始探索热点论文自身的特征，但在影响因素等规律方面的研究工作尚不充分。［方法/过程］本研究利用TF-IDF算法和负二项回归模型，试图探究热点论文的分布特征、影响因素、时间窗口差异和学科类型差异。［结果/结论］研究结果表明，热点论文的分布特征侧重于发达国家、知名研究机构、交叉学科和权威期刊；并且受到精炼的标题、国家间合作、研究型产出、开放获取、高影响因子期刊等因素影响；热点论文存在时间窗口效应，甚至改变了标题、摘要、开放获取等因素影响热点论文的具体轨迹；学科间差异对热点论文具有影响，在标题、摘要、科研合作、文献类型、开放获取、期刊影响因子方面均存在差异。

关键词：热点论文 分布特征 影响因素 时间窗口 学科差异 负二项回归模型
分类号：G301
DOI：10.13266/j.issn.0252-3116.2019.16.009

1 引言

2018年7月至9月，中共中央办公厅、国务院办公厅印发了《关于深化项目评审、人才评价、机构评估改革的意见》，国务院印发了《关于优化科研管理提升科研绩效若干措施的通知》，人社部办公厅和中科院办公厅印发了《关于深化自然科学研究人员职称制度改革的指导意见（征求意见稿）》，分别提出要“注重标志性成果的质量、贡献、影响，把学科领域活跃度和影响力等作为重要评价指标”，“建立以创新质量和贡献为导向的绩效评价体系”和“推行代表作制度，注重成果的质量、贡献、影响，将自然科学研究人员的代表性成果作为职称评审的重要内容”，表明我国科研管理部门已经深刻地认识到，重视学术界标志性、代表性成果的重要意义。并且，随着“双一流大学”建设的不断推进，已有众多高校或教师将ESI“高被引论文”（highly cited papers）或“热点论文”（hot papers）作为“标榜”自身学术影响力的方式之一。可以预见，探索“标志性、代表性成果”背景下的高被引论文或热点论文所凸显的特征及影响因素，将很快成为学术界关注的焦点话题。

学术论文是科研成果的重要载体之一，当前学术界在对论文的影响力进行评价时，依然广泛采用P. L. Gross[1]等在1927提出的“被引频次”指标。自J. A. Virgo[2]验证了被引频次与科研成果重要性的正相关假设之后，H. F. Moed等[3]指出，排除不正当或负面引用的情形，被引频次越高，往往也代表学术论文价值越大。“某一特定出版物被引用的频次越高，它对科学进步的重要性就越大”的前提是规范引用[4]，这一前提不仅为被引频次应用于研究评价中奠定了基础[5]，也表明其被运用到研究某领域具备历史根源[6]。因此，被引频次可以用来评价科研成果的重要性，反映学术论文的科研共同体认同价值[7-8]。虽然有学者证实了“睡美人文献”[9]的存在，但其依然是长引文时间窗口下基于被引频次的评价。将被引频次用于科研评价是否存在“固有缺陷”[10]，学术界虽有争议，但不可否认的是，被引频次依然是一种快速有效的评价方法，逐渐

* 本文系中国科学院发展规划局咨询课题“2018中国科学院8+2融合科学领域竞争优势国际比较的理论和方法研究”（项目编号：GHJ-ZLZX-2018-32-2）研究成果之一。

作者简介：宋超（ORCID：0000-0002-8668-6888），博士研究生；陈悦（ORCID：0000-0001-5272-9459），教授，博士，博士生导师，通讯作者，E-mail：chenyuedlut@163.com；汪玲（ORCID：0000-0002-5210-7745），硕士研究生；左佳（ORCID：0000-0002-5350-2456），硕士研究生；刘则渊（ORCID：0000-0002-9206-7207），教授，博士生导师。

收稿日期：2018-11-10 修回日期：2019-03-08 本文起止页码：84-94 本文责任编辑：王传清

成为科学计量学最常用的评价指标之一,并推广到科技政策、学科发展、图书期刊、科研人员等的评价研究[5],从而得到学术界的普遍认可。基于被引频次指标下的“高被引论文”和“热点论文”,可以在一定程度上衡量标志性、代表性成果。高被引论文和热点论文的定义,来自于科睿唯安官方网站[11],高被引论文指在 10 年内发表的论文且被引用数量处于该研究领域(research field)全球前 1% 之列,热点论文则是近两年内发表的且在近两个月内被引用次数进入该研究领域全球前 0.1% 之列的论文,都反映了所属领域中具有突破性、最有影响力的研究工作。高被引论文引文窗口较长,体现被引频次的累积过程。相比之下,热点论文则是科学研究的最新发现和研究动向,具有风向标的作用,反映了近两年内比较受关注的重要研究[12]。因此,使用热点论文衡量学者短期内具有突破意义的标志性和代表性成果,可能更具前瞻性。学术界关于高被引论文的研究工作已是屡见不鲜,然而基于热点论文的分析尚不多见。因此,探究热点论文为何热?研究热点论文的分布特征、影响因素、时间窗口与学科间差异等,具有重要的理论指导意义。

2 研究回顾与述评

2.1 研究回顾

2.1.1 以热点论文为主题的研究 关于以热点论文为主题的研究工作,除论证了其在被引方面的网络影响,还涉及热点论文的文献类型、时间窗口、学科以及期刊、国家等方面的差异。热点论文通常表现为“被引用的速度非常快”[13]。早前的研究成果显示,在网络分析中,将一篇论文视作一个节点,因此同时诞生的节点有平等的机会被连接,而关于科学论文吸引力的实证研究结论表明,热点论文可以比其他同期论文获得更多的链接或引用[14-15]。后来有学者使用“节点流行度”指标,用来刻画节点在区域中影响大小,得出区域影响较大的节点吸引连接的机会也较大,其他节点也更喜欢连接到“流行的节点”,证实了与其他同时发表的论文相比,那些比较受关注论文的被引用次数更多,并且后续的研究产出更喜欢引用的结论[16]。文献类型在热点论文方面也具有差异,相比研究类文献而言,假如综述类文献更易成为热点论文,可能表明高质量的原创性研究工作较少[17]。时间窗口的影响也不可忽略,并因学科不同而迥异。一份出版物的真正影响只能在较长时间后才能确定,如生物医学领域和多学科科学多体现为 3 年的时间窗,而对于人文和数学来说,7 年的时间窗则比较合适[18]。甚至有学者研究发现,由于物理化学领域所使用的实验方法非常专业,涉及的理论异常复杂,在短期内难以衡量该领域出版物的真正影响,并且其所提出的概念需要时间被学术界“欣赏”,因此该领域中如果有研究产出能迅速得到认可,应该得益于该领域中大量活跃的研究人员[19]。此外,相比之下,部分学科(如生物化学、分子生物学、免疫学和细胞生物学)、期刊(如 PRL、PNAS)和国家(如美国),则可能贡献更多的“hot papers”[20]。

2.1.2 热点论文影响因素方面的研究 关于热点论文影响因素的研究尚不多见,但基于被引频次影响因素的文献则可以为本研究提供参考。学术界大多在讨论论文被引频次的影响因素时,是从论文自身和外部信息来考虑的,本研究将其称之为内在因素和外在因素。前者是指论文自身所表达的信息,如标题、摘要、合作(作者、国家、机构)、文献类型、是否获得基金资助、参考文献、文章长度、发文年份等,而后者则是指刊载期刊影响因子、是否可以开放获取、文献级别使用量[21]以及学科数量等外在信息。关于上述影响因素方面的研究工作已比较常见,但研究结论迥异。

内在影响因素方面,如标题对被引频次的影响,剔除期刊编辑部对标题长度限制因素外[22],大致有 3 种研究结论:正相关[23-24],即标题越长被引频次也就越多;无关[25],即标题长度与被引频次间没有关系;负相关[26-28],即标题越短被引频次也就越多。摘要对被引频次影响的研究成果并不常见,有学者采用“弗莱士易读度”[29-30](Flesch Reading Ease)指标,对摘要可读性进行分析;同一期刊的论文,摘要中包含被频繁使用词语的论文得到的评价更高,即摘要越短、使用更常用词语,可能更容易阅读,从而获得更多的引用[31]。科研合作主体主要包括作者、国家和机构 3 个方面,合作可以丰富研究思路[32],经验研究证实合作可以改变科研绩效,提高科学产出的质量和影响力[33-34],但是 W. Glänzel 教授持不同观点,其认为“相当一部分国际合著论文被引绩效低于样本平均水平”[35]、“合作总能保证成功是一个神话”[36],还有学者指出合作规模与论文被引频次之间存在一定的不确定性[37],甚至无明显关联[38]。学术论文中标注的参考文献,构成了研究的知识基础,往往一篇学术论文的参考文献越多,其能够获得的被引也越多[39]。此外,一篇文章一经发表,便会成为后续研究的参考对象,所产出的文献类型在被引频次方面也应该存在显著差异,普遍认为综述类(review)论文的篇均被引频次明显高于研究类(arti-

cle)论文[40]。不过也存在相反结论，即认为研究类论文是高影响力论文的主要文献类型，具有较高的权威性和参考价值，而综述类论文次之[41]。获得基金资助，是开展科学研究的重要保障，可以改善研究条件、吸纳研究人员，从而提高研究产出和成果质量，即得到资助的论文成果，在发表后的被引用次数会更多[42]。全文承载了文章完整信息，论文长度即页码数[43]是影响被引频次最重要的因素，但是也有学者指出论文长度与被引频次并无关系[44]。文献平均被引用半衰期的存在，早已是不争的事实，大致约为5.6年[45]，即不同引证时间窗口[46]也会影响被引频次。

外在影响因素方面，如期刊因素对论文被引频次影响也不尽相同，通常认为刊发在高影响因子期刊上的学术论文更容易得到关注，更有可能成为高被引论文[47]。但也有学者发现大多数情况下期刊影响因子及其年度变化对被引频次并未产生直接的影响[48]。还有学者认为利用期刊影响因子评判文章是"本末倒置"之举[49]。另外，开放获取出版模式的发展[50]、不同学科间差异[51]等因素也会影响被引频次。近年来，有学者逐渐将基于索引数据库的"文献级别用量"(使用次数)指标纳入被引频次方面的研究，甚至认为可以将使用次数作为早期表征学术质量的又一指标[52]，并且指出发文在一定时间内较高的使用次数可以预示一段时期后的高被引[53-54]。学术界已敏锐地关注到文献使用次数和被引频次之间的关系，但研究结论各异：两者间呈现正相关关系[55-56]、相关关系较弱[27,57]、负相关关系[58]。虽然较早发表的高被引论文使用次数较大，但读者们更倾向于使用较新的文献[59]。另外，不同学科领域的读者数量、文献类型等存在的差异[55]，也会影响被引频次。

2.2 研究述评

通过对相关研究成果进行回顾，可以清晰地发现，目前鲜有研究利用定量的方法对热点论文的分布特征、影响因素等问题做出回答。此外，纵观基于被引频次影响因素方面的研究成果，多数是基于固定时间窗口或单一学科展开的，发文年份和学科间的差异使得跨学科、时间序列上的比较工作变得异常困难，研究结论也迥异。本研究认为，时间窗口和学科间差异可能会使得其他因素作用于被引频次的影响机制发生改变，即将时间、学科因素纳入分析过程时，产生一种调节作用，从而改变其他因素对被引频次的影响。因此本研究将重点探讨热点论文的分布特征、影响因素，并对不同时间窗口和学科间差异做出比较分析。

3 研究问题与研究设计

3.1 研究问题

热点论文能够快速地从海量的科学文献中脱颖而出，其背后可能具备一定的科学计量学规律，本文试图对此展开研究工作，即归纳热点论文国家、机构、学科、期刊分布特征，探寻作用于热点论文的内在和外在影响因素，并对不同时间窗口、不同学科类型下的差异做出分析(本研究的逻辑框架见图1)。

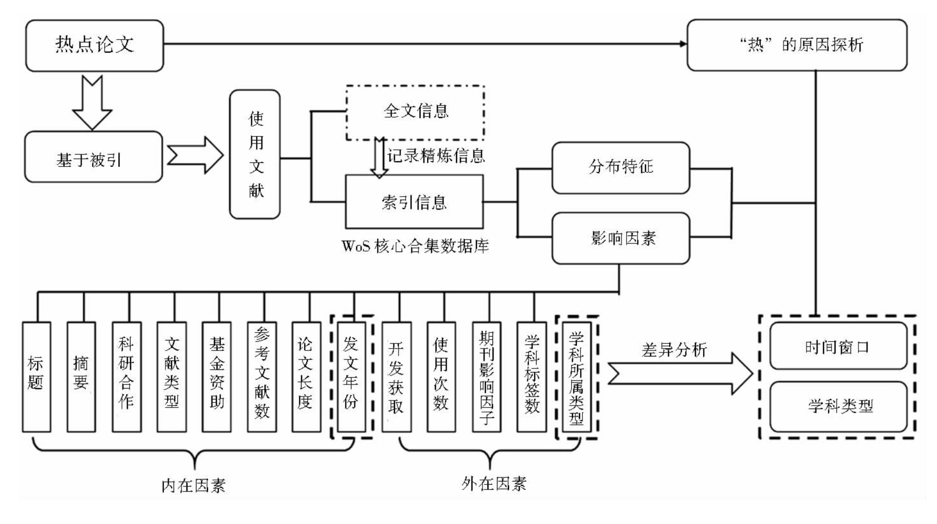


图1 研究逻辑框架

3.2 研究设计

3.2.1 数据来源 本研究基于ESI下的热点论文数据，通过Web of Science(简称WoS)核心合集数据库获取，区别于Scopus数据库"TOP25 Hottest Articles"[60]和www.Altmetric.com的"Top 100"数据[61]。由于WoS

核心合集数据库中只记录近两年的热点论文,因此将发文时间限定在 2016 年 – 2018 年,文献类型限定为"Article"和"Review",不限定学科领域,检索时间为:2018 年 9 月 2 日。检索共获得 2 959 篇热点论文,剔除字段不完整数据,保留 2 798 篇文献作为研究样本。

3.2.2 变量选取与定义 本研究基于文献内在和外在因素选取研究变量(见表 1),其中因变量为被引频次,自变量为标题实词长度、标题区别词长度、摘要实词长度、摘要区别词长度、作者合作规模、国家合作规模、机构合作规模、文献类型、基金资助、引用参考文献数、论文长度、发文年份、开放获取情况、论文使用次数、学科数、学科所属类型、期刊影响因子。对部分研究变量的处理过程做出说明:标题、摘要实词长度和区别词长度。首先利用自然语言处理技术将文章标题、摘要进行分词和停用词处理,剔除无实际意义的虚词,保留具有实际含义的词汇,这些实词能够在较大程度上表征主题,计算实词长度;然后利用 TF-IDF 算法,提取出每篇论文标题、摘要中的具有区别性的词汇,所获得的词汇长度代表每篇论文标题、摘要的区别词长度。机构合作按照作者所属一级机构计算,如一级机构内部单位合作,只记一个机构。依据钱学森的科学技术体系理论,所有的科学门类都可以划分为基础科学、技术科学和工程技术三个层次,本文按照刘则渊教授划分标准[62],将检索到的数据集中 2 637 篇自然科学类热点论文分为以上三大层次,需要说明的是,由于通常所说的技术科学都是指自然科学领域,因而本文对数据集中余下 161 篇社会科学类热点论文没有进行三个层次的细化,以此确定所属学科类型。按照 ISI Web of Knowledge 平台的《期刊引证报告》(Journal Citation Reports)所公布的数据,获取刊载热点论文期刊上年度影响因子。

表 1 各变量选取与定义

序号	变量名称	变量定义	变量符号
1	被引频次	WoS 核心合集被引频次	tc
2	标题实词长度	利用自然语言处理技术得到标题实词长度	ti_num
3	标题区别词长度	基于 TF-IDF 算法计算标题区别词长度	ti_tfidf
4	摘要实词长度	利用自然语言处理技术得到摘要实词长度	ab_num
5	摘要区别词长度	基于 TF-IDF 算法计算摘要区别词长度	ab_tfidf
6	作者合作规模	文章作者数量	co_au
7	国家合作规模	文章作者所属国家数量	co_country
8	机构合作规模	文章作者所属一级机构数量	co_organization
9	文献类型	1 = "Article";0 = "Review"	dt
10	基金资助	1 = "获得基金资助";0 = "未获得基金资助"	fu
11	引用参考文献数	文章所引用参考文献数量	nr
12	论文长度	文章总页数	pg
13	发文年份	py_2016、py_2017、py_2018 分别代表 2016、2017、2018 年发表文章	py
14	开放获取情况	1 = "可以开放获取";0 = "不可以开放获取"	oa
15	论文使用次数	2013 年至今文章被使用次数	usage
16	学科数	WoS 数据库所标注的 WC 类别数量	wc_num
17	学科所属类型	wc_basic、wc_technical、wc_engineering、wc_social 分别代表基础科学、技术科学、工程技术和社会科学	wc_subject
18	期刊影响因子	发文年份前一年 Journal Citation Reports 对应的期刊影响因子	so_factor

3.2.3 模型构建 为了研究热点论文影响因素,结合本研究所选择的指标,计量模型构建如下:

$$tc_i = f(\alpha_1 \times ti_num_i,\ \alpha_2 \times ti_tfidf_i,\ \alpha_3 \times ab_num_i,\ \alpha_4 \times ab_tfidf_i,\ \alpha_5 \times co_au_i,\ \alpha_6 \times co_country_i,\ \alpha_7 \times co_orgaization_i,\ \alpha_8 \times dt_i,\ \alpha_9 \times fu_i,\ \alpha_{10} \times nr_i,\ \alpha_{11} \times pg_i,\ \alpha_{12} \times py_i,\ \alpha_{13} \times oa_i,\ \alpha_{14} \times usage_i,\ \alpha_{15} \times wc_num_i,\ \alpha_{16} \times wc_subject_i,\ \alpha_{17} \times so_factor_i)$$

由于因变量被引频次是一个非负整数的计数变量,且不符合正态分布,呈现离散分布的特点。对此,常常选择泊松回归进行拟合,并且要求其方差和均值必须相等。然而,本研究中由于样本的方差大于均值,两者不相等,可能存在过度离散的情形,不符合泊松回归的要求[63];然后选择负二项回归,检验通过,进而选择零膨胀负二项回归,没有通过检验,故最终选择负二项回归。本研究使用 Stata14.0 对样本数据进行负二项回归分析。

4 实证结果分析

4.1 变量描述性统计和相关分析

由描述性统计指标值(见表 2)可以看出,多数变量标准差较大,不符合正态分布,呈现离散分布特点,且部分变量存在 0 值,显然不满足线性相关。对于热点论文而言,很多指标存在异常大的极差,可以预见,关于热点论文被引频次的影响因素可能不存在一般性规律。

表 2 各变量描述性统计

变量名称	样本量	均值	标准差	极小值	极大值
tc	2 798	90.81	167.1	2	4483
ti_num	2 798	9.360	3.800	1	33
ti_tfidf	2 798	1.480	1.560	0	11
ab_num	2 798	130.2	59.99	17	561
ab_tfidf	2 798	51.61	19.13	6	166
co_au	2 798	26.82	164.7	1	3614
co_country	2 798	3.360	6.840	1	105
co_organization	2 798	9.730	35.71	1	577
nr	2 798	106.4	235.5	0	8409
pg	2 798	18.28	43.50	1	1790
usage	2 798	136.1	218.1	0	2784
wc_num	2 798	1.670	1.020	1	5
so_factor	2 798	17.48	18.96	0.290	187.0

注:虚拟变量未纳入分析(本节下同)

由于变量间不满足线性相关,故使用 Spearman 相关系数对变量间进行检验,检验结果显示(见表 3),绝大多数变量通过了相关性系数显著性检验,表明所选取的变量具有一定的影响力。相关性系数处于合理范围,且方差膨胀因子 VIF 均小于 10,因此并不认为模型存在严格意义上的多重共线性,可以进行回归分析。

当样本数据量较大时,绘制热点论文被引频次与各影响因素间的散点图呈现过于密集的特征,图形可解读性不佳,因此,为克服大样本数据散点图过于拥挤的问题,本研究绘制二进制散点图(Binned Scatter-plots)[64],即将 X 轴变量分成数量相等的组,计算 X 轴和 Y 轴组内变量的均值,进而绘制其二进制散点图和总体趋势线,以更为清晰地呈现变量间的关系(见图 2),印证前面所提出的被引频次与各影响因素之间可能不存在一般性规律的经验判断。

4.2 热点论文分布特征分析

图 3 分别呈现了热点论文国家、机构、学科、期刊分布特征(排名前 10 位)。其中国家分布主要集中在发达国家,美国的确向世界贡献了最多的热点论文,值得一提的是,中国在这一方面已跃居世界第 2 位,但美国依然是中国的近两倍。大学是热点论文的主要产生机构,美国斯坦福大学位列第 1,中国科学院则位居热点论文第 2 大诞生机构,与斯坦福大学分庭抗礼。热点论文的学科分布集中在多学科科学、化学(多学科)、医学等交叉学科领域。在期刊分布方面,产生热点论文前 10 位的期刊集中于学术界顶级期刊,也印证了前述高影响力期刊更容易产生热点论文的观点。

表 3 变量间相关性分析

	tc	ti_num	ti_tfidf	ab_num	ab_tfidf	co_au	co_country	co_organization	nr	pg	usage	wc_num	so_factor
tc	1												
ti_num	-0.138***	1											
ti_tfidf	0.111***	0.440***	1										
ab_num	0.002	0.264***	0.156***	1									
ab_tfidf	0.099***	0.222***	0.233***	0.873***	1								
co_au	0.230***	0.179***	0.112***	0.313***	0.260***	1							
co_country	0.167***	0.015	0.041**	0.162***	0.148***	0.522***	1						
co_organization	0.193***	0.072***	0.061***	0.272***	0.227***	0.736***	0.748***	1					
nr	0.057***	-0.241***	-0.056***	-0.083***	0.024	-0.234***	-0.01	-0.119***	1				
pg	0.080***	-0.160***	-0.026	0.143***	0.174***	-0.057***	0.087***	0.059***	0.615***	1			
usage	0.536***	-0.118***	0.120***	-0.146***	0.009	0.004	-0.042**	-0.112***	0.324***	0.084***	1		
wc_num	-0.151***	0.115***	0.053***	-0.114***	-0.066***	-0.189***	-0.120***	-0.200***	0.065***	-0.002	0.065***	1	
so_factor	0.466***	-0.165***	0.075***	0.087***	0.136***	0.355***	0.166***	0.256***	-0.018	-0.012	0.374***	-0.297***	1

注:* $p \leqslant 0.10$, ** $p \leqslant 0.05$, *** $p \leqslant 0.01$

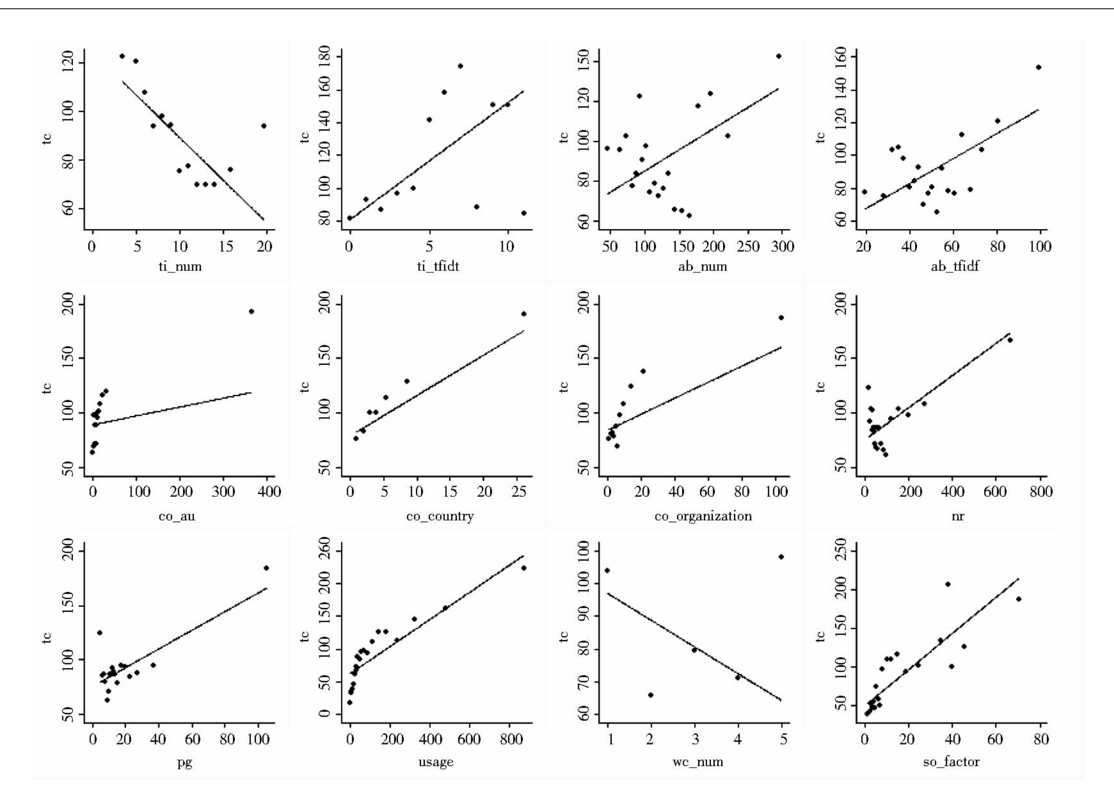


图 2　被引频次与影响因素间二进制散点图

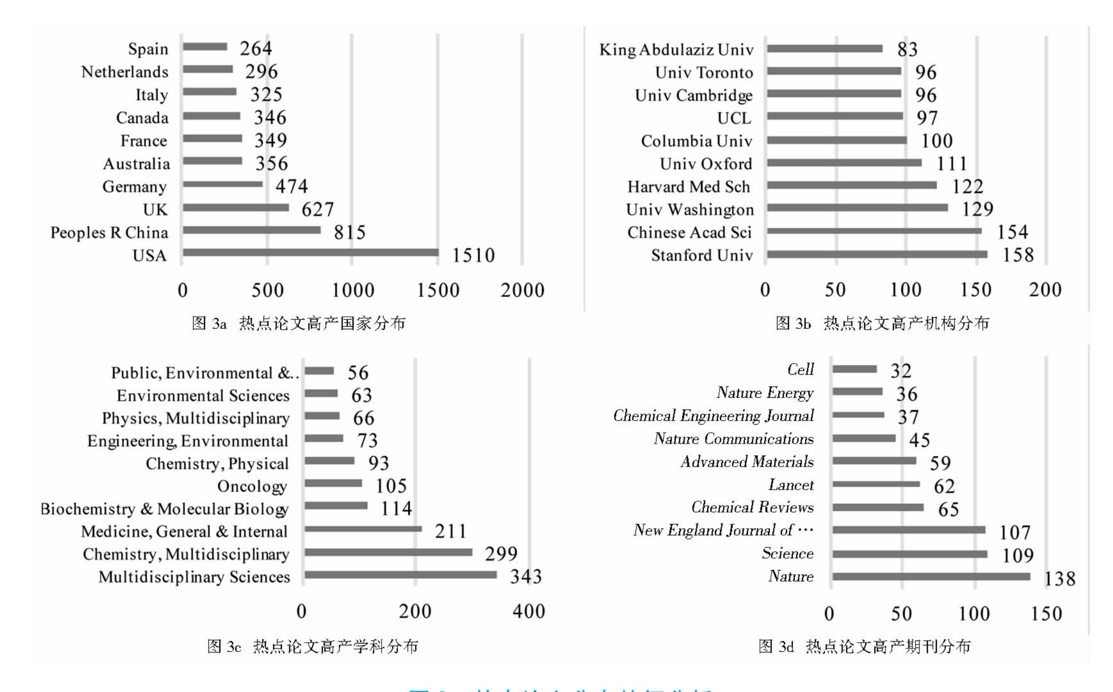


图 3a　热点论文高产国家分布

图 3b　热点论文高产机构分布

图 3c　热点论文高产学科分布

图 3d　热点论文高产期刊分布

图 3　热点论文分布特征分析

4.3　热点论文影响因素分析

4.3.1　内在影响因素方面　如表 4 所示，广泛使用区别性词汇的标题要比摘要更易成为热点论文，模型(1) – 模型(4)表明，相比摘要的大段文字，标题短小精悍，往往可以迅速抓住读者眼球，即精炼且区别于大众研究的标题，使得文章发表之后能够获得更多的被引。跨越国界的科研合作往往能够形成高质量成果，模型(1) – 模型(4)表明，作者合作、机构合作在热点论文方面影响不显著，而国家间合作具有明显的正向拉力，通常可以提高科研产出的质量，诱发研究热点。研究类文献表征所产出的热点论文具备较高的学术价值，相比综述类文献，热点论文主要体现为研究类文献，彰显研究成果的创新性。热点论文里的基金资助论文获得更高被引的可能性反而较低，模型(3)表明

基金资助的论文对其获得高被引具有负向影响，可能与文章在短期引文窗口背景下无法获得长远评价有关。广泛的知识基础对热点论文的影响关系不明显，参考文献数量越多，可以在一定意义上表明该方向上的知识基础越丰富。但模型(1)－模型(4)表明，参考文献数量对热点论文被引频次的影响不显著。论文篇幅和热点论文间影响关系存在不确定性，模型(1)－模型(4)的"文章长度"系数均未通过显著性检验，且系数有正有负，表明很难有适用于任何场合的结论。热点论文依然存在时间窗口的规律，模型(2)、模型(4)表明，2016 年、2017 年均比 2018 年的热点论文更易获得高被引，且 2016 年显著易于 2017 年，即热点论文也受时间窗口规律影响。

4.3.2　外在影响因素方面　如表 4 所示，开放获取成为催生热点论文高被引的非常显著因素，并且使用次数也构成正向影响，即一篇获得高被引的论文，一定要被读者广泛地使用和阅读，而开放获取则为之提供了便捷条件，模型(1)－模型(4)均证实了上述观点。被赋予更多学科标签的期刊，往往不利于热点论文继续获得高被引，模型(1)、模型(3)表明，学科数对于热点论文获得更高被引具有负向影响，当一本期刊涉猎过多学科时，可能因不够专业而难以获得高质量论文。技术科学领域通常可以使热点论文获得更多引用，相比社会科学，不同学科领域中，技术科学最容易促使热点论文继续获得高被引，而基础科学和工程技术间略有差异，凸显技术科学的学科影响力。高影响因子期刊更容易推动热点论文被引用，往往高影响力期刊备受学术界关注，具有风向标的作用。因此发表在高影响力期刊上的论文，也极易在整个学术界形成"追随效应"，从而加速热点论文获得高被引。

4.4　热点论文时间窗口差异分析

文献回顾和前述研究结论均表明时间窗口对热点论文具有明显影响。通过跨年份比较，可以考察哪些因素对热点论文的影响会随着时间变化而凸显或消亡(见表 5)。研究表明：越早发表的文章，其精炼的标题往往成为比摘要更能推动被引的因素，而新近发表的文章，使用有区别于其他研究的词汇撰写摘要，则显得尤为重要；科研合作作用于热点论文方面的影响不明显，国家合作可能有一定影响，但作者合作和机构合作并无显著影响；文献类型和是否收到基金资助、参考文献数量、文章长度等因素在时间窗口上的规律不明显；开放获取对推动发表较早的热点论文继续获得高被引的贡献更大，而使用量指标则对新近发表论文的作用

表 4　热点论文影响因素模型分析

	模型(1)	模型(2)	模型(3)	模型(4)
	tc	tc	tc	tc
ti_num	−0.032 8***	−0.023 5***	−0.035 4***	−0.025 9***
	(−4.54)	(−4.06)	(−4.93)	(−4.53)
ti_tfidf	0.070 8***	0.066 4***	0.068 0***	0.065 5***
	(4.89)	(5.58)	(4.69)	(5.57)
ab_num	−0.000 445	−0.001 52*	−0.000 559	−0.001 57*
	(−0.43)	(−1.89)	(−0.53)	(−1.91)
ab_tfidf	0.002 21	0.004 32*	0.002 21	0.004 34*
	(0.77)	(1.92)	(0.79)	(1.95)
co_au	0.000 266	0.000 334	0.000 250	0.000 305
	(0.83)	(1.17)	(0.77)	(1.08)
co_country	0.018 0**	0.017 9**	0.015 4*	0.015 8**
	(2.03)	(2.33)	(1.75)	(2.08)
co_organization	−0.000 520	−0.000 871	−0.000 121	−0.000 499
	(−0.27)	(−0.54)	(−0.07)	(−0.32)
dt	0.184 * *	0.201***	0.208***	0.216***
	(2.40)	(3.71)	(2.97)	(4.20)
fu	−0.082 2	−0.014 6	−0.129 * *	−0.067 7
	(−1.24)	(−0.30)	(−1.97)	(−1.42)
nr	0.000 129	0.000 262	7.17e−05	0.000 212
	(0.47)	(1.26)	(0.27)	(1.04)
pg	0.000 696	−6.12e−05	0.001 04	0.000 221
	(0.54)	(−0.06)	(0.83)	(0.23)
py_2016		1.901***		1.889***
		(32.01)		(32.47)
py_2017		1.168***		1.152***
		(23.18)		(23.23)
oa	0.491***	0.290***	0.486***	0.281***
	(8.67)	(6.64)	(8.69)	(6.57)
usage	0.001 85***	0.001 16***	0.001 90***	0.001 18***
	(12.35)	(11.66)	(12.81)	(12.18)
wc_num	−0.052 6**	−0.029 2	−0.053 7**	−0.031 6
	(−2.04)	(−1.48)	(−2.12)	(−1.63)
wc_basic			0.548***	0.525***
			(5.41)	(7.22)
wc_technical			0.608***	0.546***
			(5.85)	(7.45)
wc_engineering			0.554***	0.484***
			(4.89)	(6.08)
so_factor	0.013 5***	0.013 1***	0.012 6***	0.012 3***
	(10.70)	(12.65)	(9.73)	(11.64)
_cons	3.708***	2.387***	3.243***	1.978***
	(33.12)	(27.22)	(26.17)	(20.75)
LL	−14 809.66	−14 115.04	−14 781.595	−14 080.364
N	2 798	2 798	2 798	2 798

注：* $p \le 0.10$，** $p \le 0.05$，*** $p \le 0.01$；括号内为 Z 统计量值

更加明显；被赋予学科标签数量多的期刊，其对热点论文的影响也不稳定；此外，相比当年发表和前年发表，上年发表的热点论文在学科分类方面所体现的差异和影响更加突出，即发表后第二年更易促使热点论文获得高被引；期刊影响因子始终是正向影响，但是时间窗口规律并不明显。

表 5　热点论文时间窗口差异分析

	模型(5)	模型(6)	模型(7)
	tc	tc	tc
ti_num	－0.032 5 ***	－0.026 1 ***	－0.004 75
	(－3.32)	(－3.79)	(－0.35)
ti_tfidf	0.062 0 ***	0.072 9 ***	0.067 6 **
	(2.99)	(4.69)	(2.14)
ab_num	0.002 24	－0.002 57 ***	－0.007 53 ***
	(1.46)	(－3.76)	(－4.26)
ab_tfidf	－0.006 56	0.006 91 ***	0.017 8 ***
	(－1.53)	(3.24)	(2.98)
co_au	0.000 703	0.000 493	－1.94e－05
	(1.02)	(1.26)	(－0.04)
co_country	0.009 72	0.019 0 **	0.027 3
	(0.66)	(2.08)	(1.11)
co_organization	0.000 269	－0.002 60	－0.004 73
	(0.10)	(－1.13)	(－0.52)
dt	0.066 1	0.260 ***	0.231
	(0.62)	(4.94)	(1.56)
fu	－0.089 2	－0.104 **	0.052 9
	(－0.85)	(－1.97)	(0.53)
nr	2.49e－05	0.000 256	－0.000 342
	(0.06)	(1.11)	(－0.39)
pg	0.000 975	－0.000 225	0.004 56
	(0.48)	(－0.17)	(1.13)
oa	0.414 ***	0.248 ***	0.080 0
	(4.82)	(5.29)	(0.71)
usage	0.000 919 ***	0.001 35 ***	0.002 02 ***
	(7.39)	(11.64)	(2.99)
wc_num	0.044 0	－0.064 3 ***	－0.056 4
	(0.97)	(－3.25)	(－1.33)
wc_basic	0.439 ***	0.634 ***	0.361 ***
	(3.06)	(6.54)	(2.91)
wc_technical	0.448 ***	0.676 ***	0.408 ***
	(3.13)	(6.98)	(3.24)
wc_engineering	0.395 ***	0.545 ***	0.367 **
	(2.66)	(5.13)	(2.25)
so_factor	0.014 5 ***	0.010 7 ***	0.017 9 ***
	(7.42)	(8.96)	(6.63)
_cons	4.042 ***	3.086 ***	1.839 ***
	(22.99)	(25.07)	(9.29)
LL	－4 845.504 9	－7 501.150 8	－1 676.151 9
N	831	1500	467

注：模型(5)(6)(7)分别为 2016、2017、2018 年(py_2016、py_2017、py_2018)分析结果；* $p\leq0.10$，** $p\leq0.05$，*** $p\leq0.01$；括号内为 Z 统计量值

4.5　热点论文学科类型差异分析

学科类型差异依然是科学计量学领域不可忽视的问题，对热点论文也具有规律性影响。通过跨学科比较，可以考察各影响因素在各不同类型学科间的差异(见表 6)。即：基础科学、技术科学论文标题对热点论文获得高被引的作用大于摘要，而社会科学热点论文则比较得益于摘要的影响，工程技术热点论文受标题、摘要影响方面均不显著；基础科学领域热点论文的科研合作作用均不显著，而技术科学具有较少的作者和机构合作、较多的国家合作，工程技术体现较多的机构合作、较少的作者合作，社会科学则具有较少国际合作的特点；研究类文献在基础科学、技术科学领域更易受到关注，而在工程技术和社会科学领域则不明显；热点论文是否受到基金资助，在学科差异方面也不显著；与前述分析类似，参考文献数量对各领域热点论文均不起显著影响；文章长度对技术科学的影响较大，对其他学科影响不显著；时间窗口对所有学科热点论文的影响均十分明显，且越早发表的热点论文，越易获得被引，这与前述研究结论相同；开放获取、期刊影响因子对于基础科学、技术科学、社会科学热点论文具有正向作用，工程技术领域不显著；使用量对各学科均有显著影响。

5　结论和讨论

(1)热点论文普遍具有精炼的标题、注重国家间合作、多数属于研究型产出、可以开放获取、刊载在高影响因子期刊上等特征。发达国家的确贡献了众多热点论文，中国也具有重要影响力。热点论文往往聚焦于多学科科学等交叉领域。

(2)热点论文同其他类型文献相似，也存在时间窗口效应，时间窗口同时作用于其他因素对热点论文产生影响。时间窗口改变了标题、摘要、开放获取等因素影响热点论文的具体轨迹，即发文较早的热点论文，标题、开放获取的作用越大，而新近发表的热点论文，摘要往往受到关注，开放获取作用并不突出。

(3)学科间差异对热点论文的影响也非常重要，技术科学领域热点论文最容易获得高被引，其次是基础科学、工程技术、社会科学。不同学科领域，在标题、摘要、科研合作、文献类型、开放获取、期刊影响力等方面均存在差异。

(4)本研究仅就热点论文自身的特征和规律进行分析，并未将热点论文与其他类型文献进行比较。热点论文里可能存在“昙花一现”式文献，因此针对热点

表 6 热点论文学科间差异分析

	模型(8) tc	模型(9) tc	模型(10) tc	模型(11) tc
ti_num	-0.026 6 ***	-0.023 4 ***	0.011 8	-0.010 5
	(-3.65)	(-2.82)	(0.86)	(-0.65)
ti_tfidf	0.054 7 ***	0.081 5 ***	-0.001 32	-0.006 83
	(2.73)	(5.25)	(-0.04)	(-0.14)
ab_num	-0.002 23 **	-0.000 812	-0.003 54	-0.008 54 ***
	(-2.17)	(-1.01)	(-1.49)	(-3.73)
ab_tfidf	0.003 50	0.003 13	0.009 38	0.019 5 ***
	(1.11)	(1.22)	(1.35)	(3.45)
co_au	-0.000 813	-0.000 408 **	-0.034 5 ***	0.029 0
	(-1.23)	(-2.41)	(-3.71)	(1.50)
co_country	-0.008 33	0.022 0 ***	-9.00e-05	-0.126 **
	(-0.35)	(2.59)	(-0.00)	(-2.34)
co_organization	0.013 0	-0.002 78 *	0.097 9 *	0.018 4
	(1.49)	(-1.80)	(1.92)	(0.56)
dt	0.209 ***	0.206 **	0.163	0.067 6
	(4.13)	(2.40)	(0.85)	(0.30)
fu	-0.101	0.003 92	-0.027 5	-0.022 8
	(-1.41)	(0.05)	(-0.26)	(-0.22)
nr	5.80e-05	0.000 168	0.000 512	-0.000 438
	(0.33)	(0.40)	(0.41)	(-0.44)
pg	0.000 496	0.006 99 **	-0.004 38	0.000 875
	(0.51)	(2.14)	(-0.42)	(0.21)
py_2016	1.842 ***	1.909 ***	1.708 ***	1.709 ***
	(21.51)	(22.00)	(9.93)	(10.90)
py_2017	1.143 ***	1.171 ***	1.042 ***	0.921 ***
	(14.85)	(15.63)	(7.14)	(7.08)
oa	0.356 ***	0.163 ***	-0.046 3	0.260 **
	(5.17)	(3.05)	(-0.29)	(2.20)
usage	0.001 10 ***	0.002 48 ***	0.000 822 ***	0.002 08 *
	(8.85)	(6.43)	(3.10)	(1.86)
so_factor	0.010 7 ***	0.013 7 ***	-0.013 8	0.046 4 ***
	(6.69)	(10.03)	(-0.32)	(3.64)
_cons	2.659 ***	2.157 ***	2.368 ***	2.088 ***
	(20.77)	(15.58)	(6.91)	(6.75)
LL	-7 016.401 1	-5 606.962 4	-713.622 98	-638.401 79
N	1 358	1 117	162	161

注:模型(8)(9)(10)(11)分别为基础科学、技术科学、工程技术、社会科学(wc_basic、wc_technical、wc_engineering、wc_social)分析结果;* $p \leq 0.10$, ** $p \leq 0.05$, *** $p \leq 0.01$;括号内为 Z 统计量值

论文影响因素的研究结论,是否在其他类型文献中具有相似或截然不同的规律,尚不可知,留待后续研究进行探索。

参考文献:

[1] GROSS P L, GROSS E M. College libraries and chemical education[J]. Science, 1927, 66(1713):385-389.

[2] VIRGO J A. A statistical procedure for evaluating the importance of scientific papers[J]. Library quarterly, 1977, 47(4):415-430.

[3] MOED H F. The impact-factors debate: the ISI's uses and limits [J]. Nature, 2002, 415(6873):731-732.

[4] BORNMANN L, ANEGÓN F D M, LEYDESDORFF L. Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega Hypothesis[J]. Plos one, 2010, 5 (10):e13327.

[5] BORNMANN L, DANIEL H. What do citation counts measure? A review of studies on citing behavior[J]. Journal of documentation, 2008, 64(1):45-80.

[6] MARX W, BORNMANN L. Change of perspective: bibliometrics from the point of view of cited references: a literature overview on approaches to the evaluation of cited references in bibliometrics [J]. Scientometrics, 2016, 109(2):1397-1415.

[7] PERITZ B C. On the objectives of citation analysis: problems of theory and method[J]. Journal of the Association for Information Science & Technology, 1992, 43(6): 448-451.

[8] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(46):16569-16572.

[9] RAAN A F J V. Sleeping beauties in science[J]. Scientometrics, 2004, 59(3):467-472.

[10] DAVIS J B. Problems in using the social sciences citation index to rank economics journals[J]. American economist, 1998, 42(2): 59-64.

[11] CLARIVATE. Analyze top research output and research fronts [EB/OL]. [2018-08-01]. https://clarivate.com/products/essential-science-indicators/.

[12] 宋敏,杜尚宇,刘多,等. 国家自然科学基金资助产出热点论文分析——基于 2013-2015 年 ESI 数据[J]. 中国科学基金, 2017, 31(5): 489-493.

[13] THOR A, BORNMANN L, MARX W, et al. Identifying single influential publications in a research field: new analysis opportunities of the CRExplorer[J]. Scientometrics, 2018, 116(1):591-608.

[14] EOM Y H, FORTUNATO S. Characterizing and modeling citation dynamics[J]. Plos one, 2011, 6(9):e24926.

[15] WANG D, BARABÁSI A L. Quantifying long-term scientific impact [J]. Science, 2013, 342(6154):127-132.

[16] ZHENG X, OUYANG Z, ZHANG P, et al. Modeling the citation network by network cosmology[J]. Plos one, 2015, 10(3): e0120687.

[17] 王侠,吕传禄,孙晓希. 基于 ESI 的国际药理毒理学研究领域热点论文产出状况的分析[J]. 药学实践杂志, 2016, 34(5): 437-440.

[18] WANG J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013, 94(3):851-872.

[19] SCHOLES G D, KAMAT P V. Hot papers in physical chemistry

[J]. The journal of physical chemistry letters, 2016, 7(2):339-340.

[20] BORNMANN L, YE A Y, YE F Y. Identifying "hot papers" and papers with "delayed recognition" in large-scale datasets by using dynamically normalized citation impact scores[J]. Scientometrics, 2018, 116(2):655-674.

[21] 孙学军. SCI 新增功能"文献级别用量指标"是个什么东东?[EB/OL]. [2018-08-01]. http://blog.sciencenet.cn/blog-41174-926981.html.

[22] LETCHFORD A, MOAT H S, PREIS T. The advantage of short paper titles [J]. Royal society open science, 2015, 2(8):150266.

[23] HABIBZADEH F, YADOLLAHIE M. Are shorter article titles more attractive for citations? cross-sectional study of 22 scientific journals[J]. Croatian medical journal, 2010, 51(2):165-170.

[24] JACQUES T S, SEBIRE N J. The impact of article titles on citation hits: an analysis of general and specialist medical journals[J]. Journal of the Royal Society of Medicine short reports, 2010, 1(1):1-5.

[25] NAIR L B, GIBBERT M. What makes a 'good' title and (how) does it matter for citations? A review and general model of article title attributes in management science[J]. Scientometrics, 2016, 107(3):1331-1359.

[26] PAIVA C E, PAIVA B S R. Articles with short titles describing the results are cited more often[J]. Clinics, 2012, 67(5):509-513.

[27] SUBOTIC S, MUKHERJEE B. Short and amusing: the relationship between title characteristics, downloads, and citations in psychology articles[J]. Journal of information science, 2014, 40(1):115-124.

[28] GNEWUCH M, WOHLRABE K. Title characteristics and citations in economics[J]. Scientometrics, 2016, 110(3):1-6.

[29] DIDEGAH F, THELWALL M. Which factors help authors produce the highest impact research? collaboration, journal and document properties[J]. Journal of informetrics, 2013, 7(4):861-873.

[30] 徐庆富,康旭东,张春博. 多期刊比较视角下的论文被引频次若干影响因素研究[J]. 情报杂志,2018,37(2): 147-153.

[31] LETCHFORD A, PREIS T, MOAT H S. The advantage of simple paper abstracts[J]. Journal of informetrics, 2015, 10(1):1-8.

[32] 邱均平,曾倩. 国际合作是否能提高科研影响力——以计算机科学为例[J]. 情报理论与实践,2013,36(10):1-5.

[33] BEAVER D D. Reflections on scientific collaboration (and its study): past, present, and future[J]. Scientometrics, 2001, 52(3):365-377.

[34] FIGG W D, DUNN L, LIEWEHR D J, et al. Scientific collaboration results in higher citation rates of published articles[J]. Pharmacotherapy the journal of human pharmacology & drug therapy, 2006, 26(6):759-767.

[35] GLÄNZEL W, SCHUBERT A. Double effort = double impact? A critical view at international co-authorship in chemistry[J]. Scientometrics, 2001, 50(2):199-214.

[36] GLÄNZEL W. Seven myths in bibliometrics about facts and fiction in quantitative science studies[J]. Collnet journal of scientometrics & information management, 2008, 2(1):9-17.

[37] 钟镇. 农业经济与政策 Web of Science 期刊论文合著规模与绩效的相关性分析[J]. 中国科技期刊研究,2014,25(12):1513-1518.

[38] BORNMANN L, SCHIER H, MARX W, et al. What factors determine citation counts of publications in chemistry besides their quality? [J]. Journal of informetrics, 2012, 6(1):11-18.

[39] WEBSTER G D, JONASON P K, SCHEMBER T O. Hot topics and popular papers in evolutionary psychology: analyses of title works and citation counts in evolution and human behavior, 1979-2008[J]. Evolutionary psychology, 2009, 7(3): 348-362.

[40] VANCLAY J K. Factors affecting citation rates in environmental science[J]. Journal of informetrics, 2013, 7(2):265-271.

[41] 付中静. 国际权威期刊非可被引文献的引证特征以及对影响因子的贡献[J]. 中国科技期刊研究, 2016, 27(3):324-329.

[42] KULKARNI A V, BUSSE J W, SHAMS I. Characteristics associated with citation rate of the medical literature[J]. Plos one, 2007, 2(5):e403.

[43] ROBSON B J, MOUSQUES A. Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts[J]. Environmental modelling & software, 2016, 75(1):94-104.

[44] HASLAM N, KOVAL P. Predicting long-term citation impact of articles in social and personality psychology[J]. Psychol rep, 2010, 106(3):891-900.

[45] SCHLOEGL C, GORRAIZ J. Comparison of citation and usage indicators: the case of oncology journals[J]. Scientometrics, 2010, 82(3):567-580.

[46] 付中静. 不同引证时间窗口论文量引关系实证研究——基于论文与期刊视角[J]. 情报杂志, 2017, 36(7):128-133.

[47] DIDEGAH F, THELWALL M. Determinants of research citation impact in nanoscience and nanotechnology[J]. Journal of the Association for Information Science & Technology, 2013, 64(5):1055-1064.

[48] FINARDI U. Correlation between journal impact factor and citation performance: an experimental study[J]. Journal of informetrics, 2013, 7(2):357-370.

[49] SEGLEN P O. Why the impact factor of journals should not be used for evaluating research[J]. British medical journal, 1997, 314(7079):498-502.

[50] WANG X, LIU C, MAO W, et al. The open access advantage considering citation, article usage and social media attention[J]. Scientometrics, 2015, 103(2):555-564.

[51] VIEIRA E S, GOMES J A N F. Citations to scientific articles: its

distribution and dependence on the article features[J]. Journal of informetrics, 2010, 4(1):1 – 13.

[52] MARTÍNEZ M A, HERRERA M, CONTRERAS E, et al. Characterizing highly cited papers in social work through H-classics[J]. Scientometrics, 2015, 102(2):1713 – 1729.

[53] JAHANDIDEH S, ABDOLMALEKI P, ASADABADI E B. Prediction of future citations of a research paper from number of its internet downloads[J]. Medical hypotheses, 2007, 69(2):458 – 459.

[54] 丁佐奇. 基于 Web of Science 的论文使用次数和被引频次的相关性分析[J]. 中国科技期刊研究, 2017, 28(12):1166 – 1170.

[55] MOED H F, HALEVI G. On full text download and citation distributions in scientific-scholarly journals[J]. Journal of the Association for Information Science & Technology, 2016, 67(2):412 – 431.

[56] BRODY T, HARNAD S, CARR L. Earlier web usage statistics as predictors of later citation impact[J]. Journal of the Association for Information Science & Technology, 2006, 57(8):1060 – 1072.

[57] GUERRERO-BOTE V P, MOYA-ANEGÓN F. Relationship between downloads and citations at journal and paper levels, and the influence of language[J]. Scientometrics, 2014, 101(2):1043 – 1065.

[58] LIPPI G, FAVALORO E J. Article downloads and citations: is there any relationship? [J]. Clinica chimica acta, 2013, 415(1):195.

[59] WANG X, FANG Z, SUN X. Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count[J]. Scientometrics, 2016, 109(2):917 – 926.

[60] 盛丽娜. Scopus 数据库眼科学期刊热点论文分析[J]. 中华医学图书情报杂志, 2014, 23(2): 60 – 62,67.

[61] 匡登辉. 从 Altmetrics 热点论文看科技期刊影响力——以 Altmetric. com top 100 论文为例[J]. 中国科技期刊研究, 2016, 27(11): 1188 – 1194.

[62] 刘则渊,陈悦,侯海燕. 技术科学前沿图谱与强国战略[M]. 北京:人民出版社, 2012.

[63] 谢锋昌,韦博成,林金官. 零过多数据的统计分析及其应用[M]. 北京:科学出版社, 2013.

[64] CHETTY R, FRIEDMAN J N, LETH-PETERSEN S, et al. Active vs. Passive decisions and crowd-out in retirement savings accounts: evidence from denmark[J]. Quarterly journal of economics, 2014, 129(3):1141 – 1219.

作者贡献说明:

宋超:论文构思、数据处理、论文撰写、论文修改;
陈悦:论文构思、论文修改;
汪玲:数据资料收集;
左佳:数据资料收集;
刘则渊:提出论文修改意见。

Analysis of Distribution Characteristics and Influencing Factors of Hot Papers
——Comment on the Differences Under Time Window and Interdisciplinary

Song Chao Chen Yue Wang Ling Zuo Jia Liu Zeyuan

Institute of Science of Science and S&T Management & WISE Lab, Dalian University of Technology, Dalian 116024

Abstract: [**Purpose/significance**] In recent years, hot papers have been paid more and more attention by academia. A few research results have begun to explore the characteristics of hot papers themselves, but the research on influencing factors and other laws is still insufficient. [**Method/process**] Based on this, this study uses TF-IDF algorithm and negative binomial regression model to explore the distribution characteristics, influencing factors, time window differences and disciplines type differences of hot papers. [**Result/conclusion**] The results show that the distribution characteristics of hot papers focus on developed countries, well-known research institutions, interdisciplinary and authoritative journals, and are influenced by refined titles, inter-country cooperation, research output, open access, high-impact factor journals and other factors; hot papers have time window effect, and even change the titles, abstracts, open access and other factors. The differences among disciplines have an impact on hot papers, and there are differences in title, abstract, scientific research cooperation, literature type, open access and journal impact factors.

Keywords: hot papers distribution characteristics influencing factors time window disciplinary differences negative binomial regression model